Efficiency and Equilibrium in Trial and Error Learning

Bary S. R. Pradelski and H. Peyton Young

University of Oxford

Abstract. In trial and error learning, agents experiment with new strategies and adopt them with a probability that depends on their realized payoffs. Such rules are completely uncoupled, that is, each agent's behaviour depends only on his own realized payoffs and not on the payoffs or actions of anyone else. We show that by modifying a trial and error learning rule proposed by Young (2009) we obtain a completely uncoupled learning process that selects a Pareto optimal equilibrium whenever a pure equilibrium exists. When a pure equilibrium does not exist, there is a simple formula that relates the long-run likelihood of each disequilibrium state to the total payoff over *all* agents and the maximum payoff gain that would result from a unilateral deviation by *some* agent. This *welfare/stability trade-off criterion* provides a novel framework for analyzing the selection of disequilibrium as well as equilibrium states in finite *n*-person games.

Acknowledgements. We thank Gabriel Kreindler for suggesting a number of improvements to an earlier draft. This research was supported by grants from the Office of Naval Research (#N00014-09-1-0751) and the Air Force Office of Scientific Research (FA9550-09-1-0538).

1. Learning equilibrium in large interactive systems

Game theory has traditionally focussed on situations that involve a small number of players. In these environments it makes sense to assume that players are aware of the structure of the game and take into account the behaviour of the others. But there are many situations involving huge numbers of players where these assumptions are not particularly persuasive. Commuters in city traffic are engaged in a game because each person's choice of route affects the driving time of many other drivers. Yet it is doubtful that anyone 'knows the game' or fully takes into account the strategies of the other players as is usually posited in game theory. Other examples include trading in large decentralized markets, the routing of data on the internet, and communication protocols for sharing information between distributed sensors in military applications.

These games pose several novel and challenging questions. In particular, can such systems equilibrate even though agents are unaware of the strategies and behaviors of most (or perhaps all) of the other agents? What kinds of adaptive learning rules make sense in such environments? How long does it take to reach equilibrium assuming it can be reached at all? And what can be said about the welfare properties of the equilibria that result from particular learning rules?

In the last few years the study of these issues has been developing rapidly among computer scientists and distributed control theorists[1]. The main focus of attention has been on congestion and potential games, since these arise quite frequently in applications. For this class of games there exist simple and intuitively appealing learning procedures that cause the system to equilibrate from any initial conditions. Notable among these procedures is logit learning, in which an agent chooses actions with log probabilities that are proportional to their payoffs. In this case equilibrium occurs at a local or global maximum of the potential function, but since the potential function need not measure the overall welfare of the agents, the equilibrium selected may be quite inefficient. This is a well-known problem in congestion games for example. The problem of inefficient equilibrium selection can be overcome by a congestion pricing scheme, but this requires some type of centralized (or at least not fully decentralized) mechanism for determining the price to charge on each route (Sandholm, 1998).

[1] See in particular Papadimitirou (2001), Roughgarden (2005), Mannor and Shamma (2007), Marden et al. (2009), Asadpour and Saberi (2009), Shah and Shin (2010).

The contribution of this paper is to demonstrate a very simple class of adaptive learning rules that select an efficient equilibrium in any game with generic payoffs that possesses at least one pure Nash equilibrium. (Moreover the genericity assumption can be substantially relaxed, as we show in theorem 1.) These learning rules are *completely uncoupled*: a player's updating procedure uses only his own realized payoffs as input, not the actions or payoffs of anyone else. Thus they can be implemented even in environments where players cannot observe anything that the other players might be doing.[2]

The family of rules we propose is closely related to the trial and error learning procedure introduced by Young (2009), and more distantly related to the learning by sampling procedure proposed by Foster and Young (2006) and extended by Germano and Lugosi (2007). However, there are several crucial differences. First, we assume that when a player experiments with a new action there is a probability of not accepting it even though its payoff exceeds his current aspiration level. Second, we assume that the probability of stopping a search is related in a specific way to the probability of experimenting. These two changes in the rule significantly sharpens its equilibrium selection properties. Equally important is the fact that we can identify the *disequilibrium* states that are selected when pure equilibria do not exist. Indeed, there is a simple and intuitively appealing formula that defines the *tradeoff between welfare and stability*: a disequilibrium state is selected that maximizes a weighted combination of: *i)* the total welfare (sum of payoffs) across *all* agents and *ii)* the payoff gain that would result from a deviation by *some* agent, where the first is weighted positively and the second negatively.

2. Overview of results

A key feature of the learning rules that we shall consider is that agents can search in one of two ways. In 'quiet' search an agent occasionally experiments with new strategies and adopts a new one with a probability that increases with the realized gain in payoff compared to his previous strategy. In 'noisy' search an agent frequently tries out new strategies and adopts a new one with a probability that increases in the realized level of payoff. These two forms of search can be associated with different 'psychological states'. A *content* agent is not strongly

[2] An *uncoupled* learning rule is an updating procedure that does not use other agents' realized payoffs (or payoff functions) as inputs, although it may depend on other agents' actions. Foster and Young (2006) introduced the term *radically uncoupled* to describe updating procedures that do not use the *payoffs or actions* of any other player as input. Following Young (2009) we shall refer to such rules as *completely uncoupled* instead of radically uncoupled.

motivated to search but occasionally does so anyway (quiet search). A *discontent* agent flails around trying out new things frequently (noisy search). In the first type of search, the probability of acceptance is determined by the change in payoff, whereas in the second situation the probability of acceptance is determined by the level of payoff. The rationale is that in noisy search an agent tries out many different strategies before settling on one of them, hence the payoff level seems more salient than the change in payoff from the previous period.[3]

A second feature of the learning process is the mechanism that triggers transitions between content and discontent states. The essential idea is that a transition from content (*c*) to discontent (*d*) occurs when an agent's realized payoff goes down for several periods in succession and he did not search during those periods. In other words, a $c \to d$ transition is triggered by a (negative) change in payoff that was not instigated by the agent, but rather by a change of strategy by someone else (whom the agent cannot necessarily see). By contrast, a $d \to c$ transition occurs when an agent tires of searching and accepts his current strategy as 'good enough'.

To illustrate these ideas in a concrete case, consider a commuter who ordinarily takes a certain route to work. The realized cost is the journey time (the negative of the payoff), which depends on the routes taken by other commuters. Suppose that congestion on her usual route worsens and the journey time increases. If this persists she may become discontent and start looking actively for a different route. We hypothesize that the probability of settling on a new route depends on how short it turns out to be. This is the logic of the $c \to d \to c$ transitions. But even if our commuter is reasonably content, she may not be completely content, and thus may occasionally try out new ways to go to work. Our hypothesis is that she adopts such 'experiments' with a probability that is monotone increasing in the improved travel time.

Our purpose is to show that, by suitably adjusting the probability of acceptance in the two cases (experimenting when content versus searching when discontent), such a learning process will select a system-wide pure Nash equilibrium (if there is one); moreover, among all such equilibria it will select one that maximizes total welfare over all agents. (In the commuting example this would be an equilibrium that minimizes total driving time.) When a

[3] Learning rules with high and low rates of exploration have been studied in a variety of settings, including computer science (Bowling and Veloso, 2002) and biology (Thuijsman et al., 1995) .

pure Nash equilibrium does not exist, the process selects a δ-equilibrium for some $\delta > 0$. Moreover there is a simple formula that relates the level of stability (the size of δ) and the total welfare of the selected state as a function of the learning parameters.

It is important to note that, in our framework, 'select' means something considerably weaker than 'almost sure convergence'. When we say that the learning process *selects* an equilibrium we mean that *the process is in an equilibrium with high probability in a high proportion of all time periods.* In other words, if one were to take a snapshot of the process at a random point in time, the probability would be high that it would be at an equilibrium state. In the context of a finite state process (which is the case here), almost sure convergence is a stronger property, because it implies that with probability one the process will be in an equilibrium state from some time onwards. Recent work has shown that it is difficult to achieve almost sure convergence to equilibrium with simple and natural uncoupled rules (Hart and Mas-Colell, 2003, 2006; Babichenko, 2009), whereas it is quite easy under the probabilistic notion described above, as we shall soon see.

3. The learning model

Let $\mathcal{G}$ be an n-person game on a finite joint action space $A = \prod_i A_i$ and let i's utility function be $u_i : A \to \mathbb{R}$. The *state* of player i is a triplet $z_i = (m_i, \bar{a}_i, \bar{u}_i)$, where $\bar{a}_i \in A_i$ is the player's *benchmark action* (the action he normally plays in this state), $\bar{u}_i$ is his *benchmark payoff* (what he currently expects to get), and m_i is his *mood* (which determines his current search behaviour). There are four distinct moods: content (*c*), discontent (*d*), watchful (c^-), and hopeful (c^+). The state of the process (at a given point in time) will be denoted by a triple $z = (m, \bar{a}, \bar{u})$ where m is an *n*-vector of the moods of the various players, $\bar{a}$ is an *n*-vector of their benchmark actions, and $\bar{u}$ is an *n*-vector of their benchmark payoffs. The process evolves in discrete time periods $t = 1, 2, 3, \ldots$, where $z(t)$ is the state at time *t*.

Figure 1 depicts the possible one-period transitions for a specific player i. Other transitions may be occurring during the same period for other players, and the combination of all of these transitions determines the transition function between states. However, we can omit specific reference to the actions and payoffs of players other than i, because the only effect they have on i's transitions is through i's realized payoffs.

We explain each of these cases in turn. A *content-content transition* occurs when player i decides to experiment with a new action. This event has probability $\varepsilon > 0$. Call the resulting payoff u_i'. If u_i' is higher than i's current benchmark payoff $\bar{u}_i$, then with probability $q = \varepsilon^{G(u_i' - \bar{u}_i)}$ he adopts the experimental action and the experienced payoff as his new benchmarks. (The actions are not shown to keep the figure uncluttered.)

The second case is a *discontent-content transition.* When a player i is discontent he chooses an action at random each period, accordingly to some fixed distribution (independent of ε) with full support on A_i. Player i's current search ends when he spontaneously accepts his current action and its associated payoff as his new benchmarks. The probability of this event is $p = \varepsilon^{F(u_i)}$.[4]

An indirect transition occurs when player i's payoff changes and he does *not* experiment. Thus the change must have been caused by some other player (whom he may not be able to observe). We call this a *passive change* in payoff. It causes player i to go on the alert: if the new payoff u_i' is higher than his current benchmark $\bar{u}_i$ his mood changes to c^+, whereas if the new payoff is lower than his current benchmark his mood changes to c^-. In the ensuing period he moves from c^- to d if his payoff stays below $\bar{u}_i$, and moves from c^+ to c if his payoff is above or equal to $\bar{u}_i$. (Note that i's payoff in the first period (u_i') may differ from his payoff in the second (u_i'') because of changes in the behaviour of other players.) Payoff reversals relative to the benchmark cause player i to flip-flop between c^- and c^+, so these transient states can persist for a number of periods.

[4] In Young (2009) this probability was assumed to be bounded away from zero and not dependent on ε. It was also assumed that the outcome of an experiment was always accepted if it resulted in an increase in payoff relative to the benchmark. It turns out that these seemingly minor differences in the set-up have major implications in the equilibrium selection properties of the learning process.

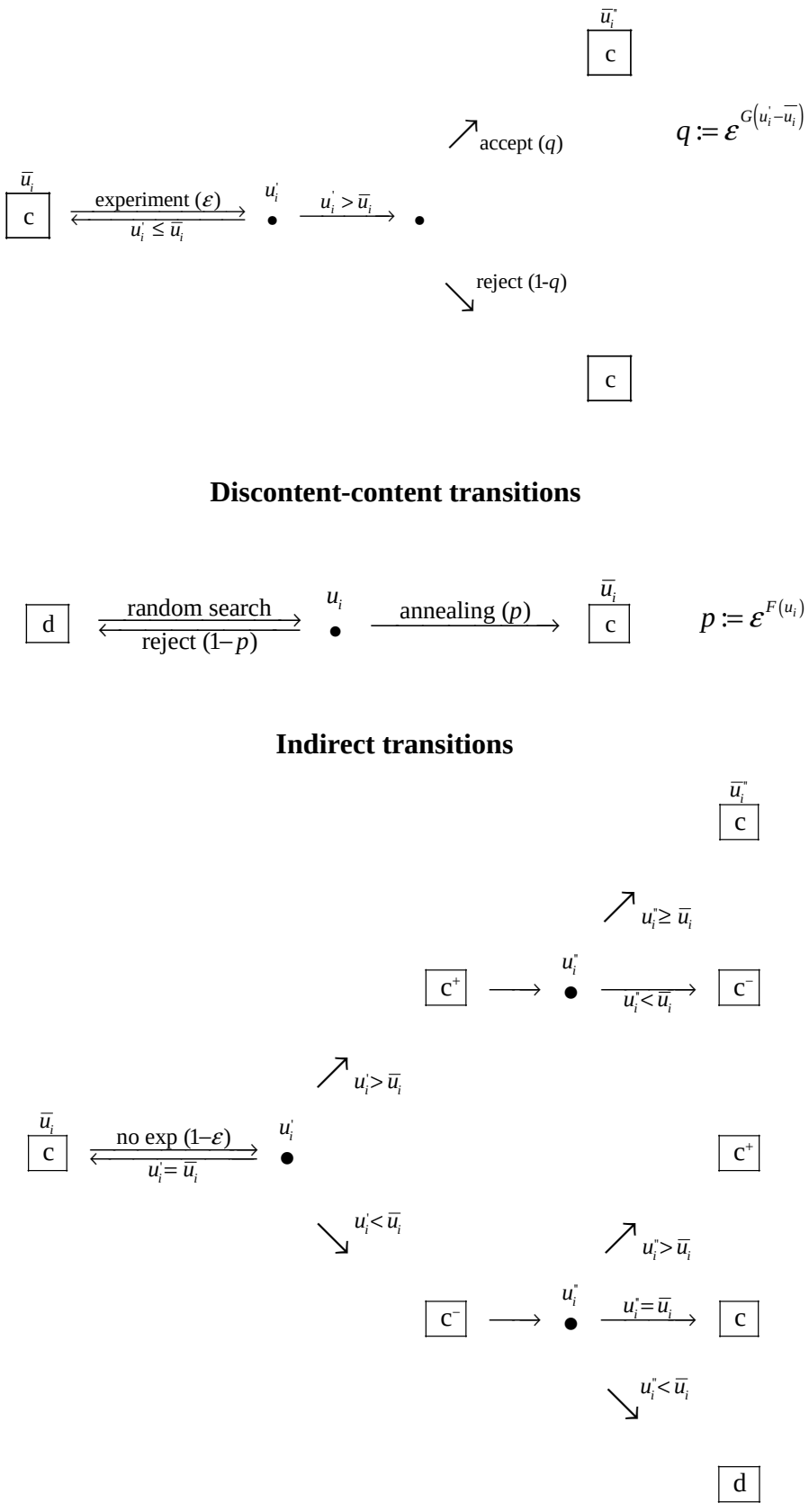

Figure 1. The structure of transitions for a given player *i*.

In summary, the parameters of the learning process consist of the *experimentation probability* ε and the two *acceptance functions* $F(u)$ and $G(\Delta u)$. Note that the domain of F is the set of payoff levels $U = \cup_i \{u_i(a) : a \in A\}$, while the domain of G is the set of nonnegative payoff differences $\Delta U = \cup_i \{[\max\{0, u_i(a_i', a_{-i}) - u_i(a_i, a_{-i})]\} : a_i, a_i' \in A_i, a_{-i} \in A_{-i}\}$. By rescaling and translating the individual payoff functions we can assume without loss of generality that $0 \le u_i(a) \le 1$ for all i and thus both U and ΔU are subsets of $[0,1]$.

For ease of exposition we shall assume that F and G are *nonnegative, strictly decreasing, linear functions*. We shall also assume, for the time being, that these acceptance functions are the same for all players. Later we shall show how the results extend to the situation where the acceptance functions differ and are non-linear. Let

$$F(u) = -f_1 \cdot u + f_2, \text{ where } f_1 > 0, \tag{1}$$

$$G(\Delta u) = -g_1 \cdot \Delta u + g_2, \text{ where } g_1 > 0. \tag{2}$$

We shall also assume that these function satisfy the following bounds

$$0 < G(\Delta u) < 1/2 \text{ and } 0 < F(u) < 1/2n. \tag{3}$$

Condition (3) implies that the acceptance probabilities are considerably larger than the probability of an experiment ($\varepsilon^p > \varepsilon$ when $p < 1$). These bounds are not best possible but they are easy to work with and involve only a modest loss in generality.

Welfare. The *welfare* of a state $z = (m, \bar{a}, \bar{u})$ is the sum of the players' payoffs from their benchmark actions:

$$W(z) = \sum_{i=1}^{n} u_i(\bar{a}). \tag{4}$$

δ-equilibrium. An action-tuple $\bar{a} \in A$ is a *δ-equilibrium* for some $\delta \ge 0$ if

$$\forall i, \ \forall a_i \in A_i, \ \ u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a}) \le \delta \ . \tag{5}$$

Stability. The (*in*)*stability* of a state $z=(m,\bar{a},\bar{u})$ is the minimum $\delta \geq 0$ such that $\bar{a}$ is a δ-*equilibrium*:

$$S(z) = \min\{\delta : \text{the benchmark actions constitute a } \delta\text{-equilibrium}\}. \qquad (6)$$

.

Stochastic stability. The set of *stochastically stable states* Z^* is the minimal subset of states such that, given any small $\alpha > 0$, there is a number $\varepsilon_\alpha > 0$ such that whenever $0 < \varepsilon \leq \varepsilon_\alpha$, $z(t) \in Z^*$ at least $1-\alpha$ of all times *t*.

Interdependence. An *n*-person game $\mathcal{G}$ on the finite action space A is *interdependent* if, for every $a \in A$ and every proper subset of players $\varnothing \subset J \subset N$, there exists some player $i \notin J$ and a choice of actions a_J' such that $u_i(a_J', a_{N-J}) \neq u_i(a_J, a_{N-J})$.

In other words, given any current choice of actions $a \in A$, any proper subset of players *J* can cause a payoff change for some player not in J by a suitable change in their actions. Note that if a game has generic payoffs (and therefore no payoff ties), this condition is automatically satisfied. However it is a much weaker condition. Consider, for example, a traffic game in which agents are free to choose any route they wish. There are many payoff ties because a local change of route by one player does not change the payoffs of players who are using completely different routes. But it satisfies the interdependence condition because a given player, or set of players, can (if they like) switch to a route that is being used by another player and thereby change his payoff.

Equilibrium state. A state z is an *equilibrium state* if all players are content, their benchmark actions constitute a Nash equilibrium and their benchmark payoffs are the utilities that result from these actions.

Theorem 1. *Let $\mathcal{G}$ be an interdependent n-person game on a finite joint action space A. Suppose that all players use trial and error learning with experimentation probability ε and acceptance functions F and G satisfying conditions (1)-(3).*

i) If $\mathcal{G}$ has at least one pure Nash equilibrium, then every stochastically stable state is an equilibrium state that maximizes $W(z)$ among all equilibrium states;

ii) If $\mathcal{G}$ has no pure Nash equilibrium, every stochastically stable state maximizes

$$f_1 W(z) - g_1 S(z) \ . \qquad (7)$$

4. Examples

Before turning to the proof we illustrate the result with two simple examples.

Example 1. Let $\mathcal{G}$ be a symmetric 2×2 coordination game with payoff matrix

	A	B
A	a,a	c,d
B	d,c	b,b

Assume that the equilibrium AA is strictly risk-dominant, that is $a-d>b-c>0$. Let us also assume that the equilibrium BB is Pareto optimal, that is, $b>a$. By theorem 1 the learning process selects BB, benchmark actions are BB a very large proportion of the time (and hence these actions are played a very large proportion of the time). This contrasts with many other adaptive learning procedures (such as logistic learning) that select the risk dominant equilibrium in 2 x 2 games (Kandori, Mailath and Rob, 1993; Young, 1993; Blume, 1993, 1995, 2003).

Example 2. Let $\mathcal{G}$ be a 2×2 game with payoff matrix

	A	B
A	$30,30$	$0,40$
B	$24,23$	$10,20$

This game has no pure Nash equilibria. The values of W and S for each of the four action-tuples are shown in Figure 2.

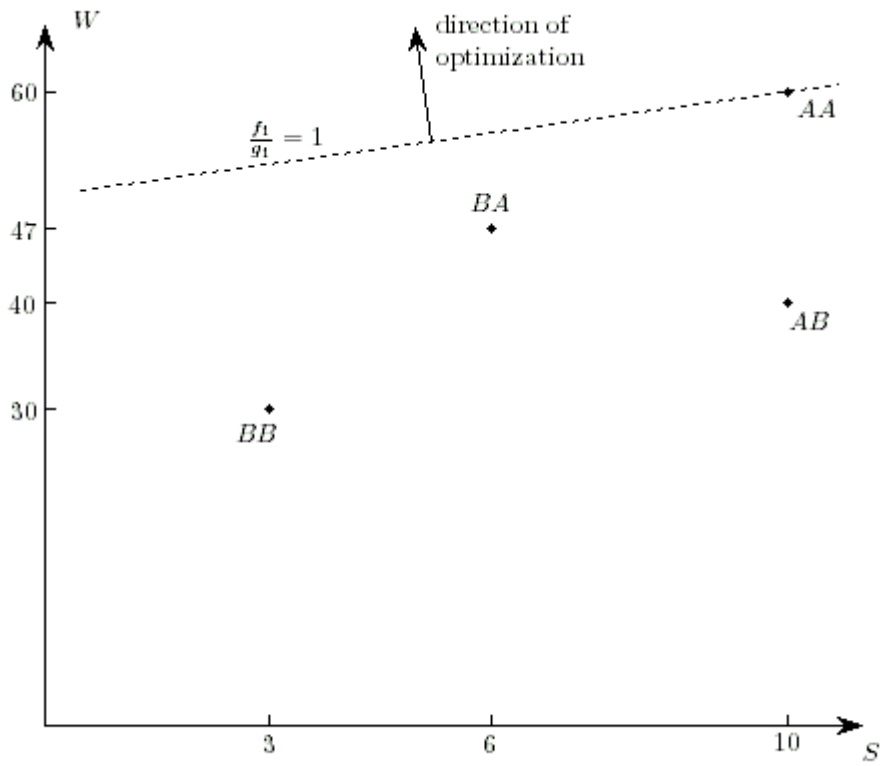

Figure 2. The tradeoff between welfare (W) and stability (S) when no equilibrium exists.

Theorem 1 implies that the learning process selects the combination that maximizes $f_1 \cdot W - g_1 \cdot S$. Figure 2 illustrates the case $f_1 / g_1 = 1$ in which AA is selected. In general the outcome depends on the ratio f_1 / g_1 as shown in Figure 3. Thus the welfare maximizing state AA is selected whenever f_1 / g_1 is sufficiently large, that is, whenever the marginal change in the rate of acceptance by a discontent player (for a given small change in experienced payoff), is sufficiently large relative to the marginal change in the rate of acceptance of an experiment (for a given small change in the gain in payoff).

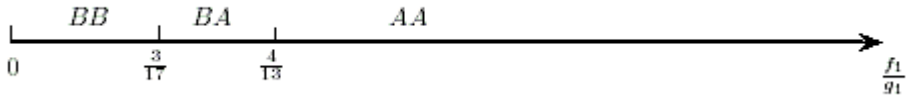

Figure 3. Stochastically stable outcomes as a function of f_1 / g_1

5. Proof of theorem 1.

Before turning to the details we give an intuitive outline of the argument. The learning process moves through a series of steps in which players experiment and adopt new actions provided they result in higher payoffs. There are two possibilities.

1. A sequence of experiments leads to increasing payoffs for some and no decrease for anyone. In this case everyone eventually becomes content with higher payoff benchmarks.

2. A sequence of experiments ends with someone's payoff going down and staying down for two periods in a row. This person becomes discontent and starts searching widely. With some probability (not dependent on ε) his searching causes other players to become discontent and they start searching. It can be shown, in fact, that the probability that a single discontent player will eventually cause all players to become discontent is bounded away from zero independently of ε.

The process only re-enters a content state once everyone has settled down again. By assumption the probability that everyone settles on a *particular* combination of actions $a \in A$ is proportional to $\prod_i \varepsilon^{F(u_i(a))}$, which is proportional to $\varepsilon^{-f_1 \sum_i u_i(a)}$. Thus high welfare states have an advantage in the sense that the process flows into them with higher probability than to other states.

However, one must also consider the probability that the process exits from any given state. Here the equilibria are advantaged because it requires at least two experiments to exit to another (nontransient) state, whereas to exit from a disequilibrium state requires at most one experiment. The essence of the proof is to show that when pure Nash equilibria exist, stability takes precedence over welfare, whereas if no pure equilibrium exists there is an explicit *tradeoff* between welfare and stability.

We now turn to the details of the argument. The trial and error learning process defines a finite Markov chain on the state space Z. For every two states $z, z' \in Z$ there is a probability (possibly zero) of transiting from z to z' in one period. We shall write this one-period transition probability as a function of the experimentation rate ε: $P_{zz'}^{\varepsilon}$. The transition is

feasible if $P^{\varepsilon}_{zz'} > 0$ whenever $\varepsilon > 0$. The *resistance* of a feasible transition $z \to z'$ is the unique real number $r(z,z') \geq 0$ such that $0 < \lim_{\varepsilon \to 0^+} P^{\varepsilon}_{zz'} / \varepsilon^{r(z,z')} < \infty$. The resistance of an infeasible transition is defined to be $r(z,z') = \infty$. If the probability of a transition is bounded away from zero when ε is small, we shall say that the transition has *zero resistance* or equivalently has probability of order $O(1)$. Similarly a transition with resistance r has probability of order $O(\varepsilon^r)$.

In general, a *recurrence class* of a finite Markov chain is a nonempty subset of states $R \subseteq Z$ such that: i) for every two distinct states $z, z' \in R$ there is a positive probability path from z to z' and a positive probability path from z' to z; ii) there is no positive probability path from any state in R to a state outside of R. The first step in the proof is to characterize the recurrence classes of the unperturbed process P^0, that is, the process when $\varepsilon = 0$. In this situation, no one experiments and no one converts from being discontent to content.

Aligned. The benchmarks in state $z = (m, \overline{a}, \overline{u})$ are *aligned* if the benchmark payoffs result from playing the benchmark actions, that is, if $\overline{u}_i = u_i(\overline{a})$ for every player i.

Notation. Let Z^0 be the subset of states in which everyone's benchmarks are aligned. Let C^0 be the subset of Z^0 in which everyone is content and let E^0 be the subset of C^0 in which the benchmark actions constitute a pure Nash equilibrium. Finally, let

$$D = \{\text{all states } z \in Z \text{ in which } \textit{every} \text{ player is discontent}\}. \qquad (8)$$

Recall that if a player is discontent, he chooses an action according to a distribution that has full support and is independent of the current benchmark actions and payoffs. Moreover, the probability of accepting the outcome of such a search depends only on its realized payoff, and the old benchmarks are discarded. Hence, for any $w \in D$ and any z, the probability of the transition $w \to z$ is independent of w. It follows that we can collapse all the states in D to a single state, which we shall continue to denote by D. Note also that $\{D\}$ is now a singleton recurrence class in the unperturbed process.

Lemma 1. The recurrence classes of the unperturbed process are D and all singletons $\{z\}$ such that $z \in C^0$.

Proof. First we shall show that these *are* recurrence classes of the unperturbed process $(\varepsilon = 0)$. Then we shall show that there are no other recurrence classes.

Suppose that $z \in C^0$, that is, the benchmarks are aligned and everyone is content. Since $\varepsilon = 0$, no one experiments and everyone replays his benchmark action next period with probability one. Hence the process remains in state z with probability one.

Next suppose that $z \in D$, that is, everyone is discontent. The probability that any given player becomes content next period is $\varepsilon^{F(\cdot)} = 0$. (Recall that $F(\cdot)$ is strictly positive.) Hence the probability is one that the process stays in D forever.

It remains to be shown that there are no other recurrence classes of the unperturbed process. We first establish the following.

Claim. Given any state z in which at least one player is discontent, there exists a sequence of transitions in the unperturbed process to the set D.

Consider a state in which some player i is discontent. By interdependence he can choose an action that alters the payoff of someone else, say j. Assume that i plays this action for two periods in a row. If j's payoff *decreases* then in two periods he will become discontent also. If j's payoff *increases* then in two periods he will become content again with a higher payoff benchmark At this point there is a positive probability that the first player, i, will revert to playing his original action for two periods. This causes player j's payoff to decrease relative to the new benchmark. Thus there is a positive probability that, in four periods or less, i's behaviour will cause j to become discontent. (The argument implicitly assumed that no one except for i and j changed action in the interim, but this event also has positive probability.) It follows that there is a series of transitions to a state where both i and j are discontent. By interdependence there are actions of i and j that cause a third player to become discontent. The argument continues in this manner until the process reaches a state where all players are discontent, which establishes the claim.

An immediate consequence is that D *is the only recurrence class that contains a discontent player*. Thus, to conclude the proof of lemma 1, it suffices to show that if z is a state in which no one is discontent, then there is a finite sequence of transitions to D or to C^0. Suppose that someone in z is in a transient mood (c^+ or c^-). Since no one is discontent and no one experiments, there is no change in the players' actions next period, and therefore no change in their realized payoffs. Hence everyone in a transient mood switches to *c* or *d* in one period. If anyone switches to *d* there is a series of transitions to D. Otherwise everyone becomes content (or already was content) and their benchmarks are now aligned, hence the process has arrived at a state in C^0. □

We now apply a theorem of Young (1993, theorem 4) which shows that the computation of the stochastically stable states can be reduced to an analysis of rooted trees on the vertex set R consisting solely of the recurrence classes. By Lemma 1, R consists of the singleton states in C^0, and also the class D. When we write $z \in R$ we shall mean that z is one of these singleton classes (or D), but we omit the singleton set notation.

Edge resistance. For every pair of *distinct* recurrence classes w and z, let $r(w \to z)$ denote the total resistance of the least-resistant path that starts in w and ends in z. We call $w \to z$ an *edge* and $r(w \to z)$ the *resistance* of the edge.

r-function.* Define the function $r^*(z)$ as follows

$$\forall z \in R, \qquad r^*(z) = \min\{r(z \to w) : w \in R - \{z\}\}. \tag{9}$$

Easy. An *easy edge* from a state z is an edge $z \to w$, $w \neq z$, such that $r(z \to w) = r^*(z)$. An *easy path* is a sequence $w^1 \to w^2 \to ... \to w^m$ in which each edge is easy and all states are distinct. An *easy tree* is a tree all of whose directed edges are easy.

The next three lemmas evaluate the function r^* on the various types of recurrence classes.

Lemma 2. $\forall e \in E^0, \qquad r^*(e) = 2$ and $e \to D$ is an easy edge.

Proof. Let $e=(m,\bar{a},\bar{u})\in E^0$, where $\bar{a}$ is a pure Nash equilibrium. Consider any outgoing edge $e\to z$ where $z\in C^0$ and $z\neq e$. Since e and z are distinct and everyone is content in both states, they must differ in their benchmark actions. Now consider any path from e to z in the full state space Z. Along any such path at least two players must experiment with new actions (an event with probability $O(\varepsilon^2)$), or one player must experiment twice in succession (also an event with probability $O(\varepsilon^2)$) in order for someone's benchmark action or payoff to change. (A single experiment is not accepted by the experimenter because $\bar{a}$ is a Nash equilibrium. Thus, although some other players may become hopeful or watchful, this is a temporary phenomenon and they revert to being content with their old benchmarks unless a second experiment occurs in the interim.) It follows that $r(e\to z)\geq 2$.

It remains to be shown that the resistance of the transition $e\to D$ is exactly two, which we shall do by constructing a particular path from e to D in the full state space. Choose some player i. By interdependence there exists an action $a_i\neq\bar{a}_i$ and a player $j\neq i$ such that $u_j(a_i,\bar{a}_{-i})\neq u_j(\bar{a})$. Let player i experiment by playing a_i twice in succession. This event has probability $O(\varepsilon^2)$, so the associated resistance is two. If $u_j(a_i,\bar{a}_{-i})>u_j(\bar{a})$, player j's mood changes to c^+ after the first experiment and to c again after the second experiment. At this point, however, j has a new higher benchmark, namely, $u_j(a_i,\bar{a}_{-i})$. Now with probability $(1-\varepsilon)^2$ player i reverts to playing $\bar{a}_i$ for the *next* two periods, which causes j to become discontent. By Lemma 2 this leads to an all-discontent state with no further resistance. The other case, namely $u_j(a_i,\bar{a}_{-i})<u_j(\bar{a})$, also leads to an all-discontent state with no further resistance. We have therefore shown that $r(e\to D)=2$, and hence that $e\to D$ is an easy edge. □

Lemma 3. $\forall z\in C^0-E^0$, $r^*(z)=1+G(S(z))$, and if $z\to z'$ is an easy edge with $z'\in C^0$, then $W(z)<W(z')$.

Proof. Let $z=(m,\bar{a},\bar{u})\in C^0-E^0$, in which case $\bar{a}$ is not a Nash equilibrium. Then there exists an agent i and an action $a_i\neq\bar{a}_i$ such that $u_i\left(a_i,\bar{a}_{-i}\right)>u_i\left(\bar{a}\right)=\bar{u}_i$. Among all such agents i and actions a_i suppose that $\Delta u_i=u_i\left(a_i,\bar{a}_{-i}\right)-u_i\left(\bar{a}\right)$ is a maximum. Let i experiment once with this action and accept the outcome of the experiment. The probability of

this event is $O(\varepsilon^{1+G(\Delta u_i)})$. If the experiment causes everyone else's payoff to stay the same or go up, and if no one experiments in the next period (an event with probability $(1-\varepsilon)^n$), then a state z' is reached after one period in which everyone is content and the benchmarks are aligned $(z' \in C^0)$ and $W(z) < W(z')$. The total resistance of this path is $r(z \to z') = 1 + G(S(z))$. The only other possibility is that the experiment causes someone else's payoff to decrease, so with no further experiments the process transits to an all-discontent state. Hence in this case we have $r(z \to D) = 1 + G(S(z))$. Hence if there is only one experiment, the resistance to exiting z is $1 + G(S(z))$. If there are two or more experiments, the resistance is at least two. By assumption, however, $G(\cdot) < 1/2$, so making two experiments involves a higher resistance than making one experiment and accepting the outcome. It follows that $r^*(z) = 1 + G(S(z))$, and that if $z \to z' \in C^0$ is an easy edge, then $W(z) < W(z')$. □

Lemma 4. $\forall z = (m, \bar{a}, \bar{u}) \in C^0$, $r(D \to z) = \sum_i F(u_i(\bar{a}))$ and $r^*(D) = \min_{a \in A} \sum_i F(u_i(a))$.

Proof. Let $D \to z = (m, \bar{a}, \bar{u}) \in C^0$. Given any $w \in D$ the probability is $O(1)$ that next period every player i chooses $\bar{a}_i$. In this case their realized payoffs are $\bar{u}_i = u_i(\bar{a})$. They all accept these actions and payoffs as their new benchmarks with probability $\prod_i \varepsilon^{F(u_i(\bar{a}))} = \varepsilon^{\sum_i F(u_i(\bar{a}))}$. Therefore $r(D \to z) \leq \sum_i F(\bar{u}_i)$.

We claim that in fact $r(D \to z) = \sum_i F(\bar{u}_i)$. Let $z = (m, \bar{a}, \bar{u}) \in C^0$ and consider a least-resistant path $w^1, w^2, ..., w^m = z$ where $w^1 \in D$. For each player i there must be some time in the sequence where i was discontent and accepted a benchmark payoff that was $\bar{u}_i$ *or less*. (There may also have been a time when i was discontent and accepted a benchmark payoff that was strictly *more* than $\bar{u}_i$, but in that case there must have been a later time at which he became discontent again. In other words eventually i accepted a benchmark payoff that was $\bar{u}_i$ or less.) The probability of such an acceptance is at most $\varepsilon^{F(\bar{u}_i)}$, because the probability of acceptance is increasing in u_i. (Recall that $F(u)$ is strictly decreasing in u.) Since this applies to every player, the total resistance of this path from D to z must be *at least*

$\sum_i F(\bar{u}_i)$. This proves that $r(D \to z) = \sum_i F(\bar{u}_i)$. The claim that $r^*(D) = \min_{a \in A} \sum_i F(u_i(a))$ follows from the fact that F is monotone decreasing. □

*w-tree***.** Identify the recurrence classes with the nodes of a graph. Given a recurrence class (node) *w*, a collection of directed edges T forms a *w-tree* if from every node $z \neq w$ there is exactly one outgoing edge in T and there is a unique directed path in T from z to *w*.

*Stochastic potential***.** The *resistance* $r(T)$ of a *w*-tree T is the sum of the resistances of its edges. The *stochastic potential* of *w* is $\rho(w) = \min\{r(T) : T \text{ is a } w\text{-tree}\}$.

In the next two lemmas we compute the stochastic potential of each type of recurrence class. From these computations theorem 1 will follow almost immediately.

Lemma 5. There exists a D-tree T_D^* that is easy.

Proof. From each node $z \neq D$ choose an easy outgoing edge $z \to \cdot$. If there are several such edges choose one that points to D, that is, choose $z \to D$ if it is easy. This implies in particular that for every $e \in E^0$ we select the edge $e \to D$. (This follows from Lemma 2.) We claim that the collection of all such edges forms a D-tree. To establish this it suffices to show that there are no cycles. Suppose by way of contradiction that $z^1 \to z^2 \to ... \to z^m \to z^1$ is a shortest cycle. This cycle cannot involve any node in E^0, because by construction the outgoing edge from any such edge points towards D, which has no outgoing edge. Therefore all $z^k \in C^0 - E^0$. Since all of the edges $z^k \to z^{k+1}$ are easy, Lemma 3 implies that $W(z^k) < W(z^{k+1})$. From this we conclude $W(z^1) < W(z^m) < W(z^1)$, which is impossible. □

Let
$$\rho^* = \rho(D) = r(T_D^*). \tag{10}$$

Lemma 6. For every $z \in C^0$ let $z \to w_z$ be the unique outgoing edge from z in T_D^* and define

$$T_z^* = T_D^* \text{ with } z \to w_z \text{ removed and } D \to z \text{ added} . \tag{11}$$

T_z^* is a z-tree of least resistance and hence

$$\rho(z) = \rho^* - r(z \to w_z) + r(D \to z). \qquad (12)$$

Proof. Plainly, the tree T_z^* defined in (11) *is* a *z*-tree. Furthermore all of its edges are easy except possibly for the edge $D \to z$. Hence it is a least-resistant *z*-tree among all *z*-trees that contain the edge $D \to z$. We shall show that in fact it minimizes resistance among all *z*-trees.

Let T_z be some *z*-tree with minimum resistance, and suppose that it does not contain the edge $D \to z$. Since it is a spanning tree it must contain some outgoing edge from D, say $D \to z'$. We can assume that $r(D \to z') < r(D \to z)$, for otherwise we could simply take out the edge $D \to z'$ and put in the edge $D \to z$ to obtain the desired result.

Let $u_1, u_2, ..., u_n$ be the benchmark payoffs in z and let $u_1', u_2', ..., u_n'$ be the benchmark payoffs in z'. By lemma 4,

$$r(D \to z) = \sum_i F(u_i) \text{ and } r(D \to z') = \sum_i F(u_i'). \qquad (13)$$

Since $r(D \to z') < r(D \to z)$ and $F(u_i)$ is monotone decreasing in u_i, we must have $\sum_i u_i < \sum_i u_i'$. Let $I = \{i : u_i' > u_i\}$. Consider the unique path in T_z that goes from z' to z, say $z' = z^1, z^2, ..., z^m = z$. Each edge $z^k \to z^{k+1}$ corresponds to a sequence of transitions in the full state space Z, and the union of all these transitions constitutes a path in Z from z' to z. Along this path, each player $i \in I$ must eventually lower his payoff benchmark because his starting benchmark u' is greater than his ending benchmark u_i. Thus at some point *i* must become discontent and adopt a new benchmark that is u_i or lower. Assume that this happens in the transition from z^{k_i} to z^{k_i+1}. Since $F(u_i)$ is strictly decreasing, the probability of adopting a benchmark that is u_i or lower is at most $\varepsilon^{F(u_i)}$. Hence the resistance of the transition from z^{k_i} to z^{k_i+1} satisfies the inequality

$$\begin{aligned} r(z^{k_i} \to z^{k_i+1}) &\geq \min_{w \in Z, w \neq z^{k_i}} r(z^{k_i} \to w) + F(u_i) \\ &= r^*(z^{k_i}) + F(u_i) = 1 + G(S(z)) + F(u_i) \end{aligned}. \qquad (14)$$

(Recall from Lemma 3 that the entire resistance along an easy path out of z^{k_i} occurs in the first step and equals $1+G(S(z))$.) It follows that the total resistance along the sequence $z' = z^1, z^2, ..., z^m = z$ satisfies

$$\sum_{k=1}^{m-1} r(z^k \to z^{k+1}) \geq \sum_{i \in I} F(u_i) + \sum_{k=1}^{m-1} r^*(z^k). \tag{15}$$

The resistance of the edge $D \to z'$ satisfies

$$r(D \to z') = \sum_i F(u_i') > \sum_{i \notin I} F(u_i') \geq \sum_{i \notin I} F(u_i). \tag{16}$$

Hence in the tree T_z the outgoing edges from the nodes $\{D, z^1, z^2, ..., z^{m-1}\}$ have a total resistance that is strictly greater than $\sum_i F(u_i) + \sum_{k=1}^{m-1} r^*(z^k)$. But in the tree T_z^* the edges from these nodes have a total resistance equal to $\sum_i F(u_i) + \sum_{k=1}^{m-1} r^*(z^k)$. Furthermore, at every other node the resistance of the outgoing edge in T_z is at least as great as it is in T_z^* (because the latter consists of easy edges). We conclude that T_z^* must minimize resistance among all z-trees , which completes the proof of lemma 6. □

To complete the proof of theorem 1, we shall first prove the following chain of inequalities:

$$\rho(e^*) \overset{(i)}{<} \rho(e) \overset{(ii)}{<} \rho(z) \overset{(iii)}{<} \rho(D) = \rho^* \tag{17}$$

with $e^* \in E^*$, $e \in E^0 - E^*$, $z \in C^0 - E^0$.

Let $e = (m, \bar{a}, \bar{u}) \in E^0$ be an equilibrium state. By construction $e \to D$ is an edge in the easy tree T_D^* (see the beginning of the proof of Lemma 5), so (12) implies that

$$\rho(e) = \rho^* - r(e \to D) + r(D \to e). \tag{18}$$

From Lemma 2 we know that $r(e \to D) = 2$. From Lemma 4 we know that

$$r(D \to e) = \sum_i F(u_i(\bar{a})) = -f_1 W(e) + nf_2 . \tag{19}$$

From this and (18) we conclude that,

$$\rho(e) = \rho^* - 2 - f_1 W(e) + nf_2 . \tag{20}$$

Hence $\rho(e^*) < \rho(e)$ whenever $W(e^*) > W(e)$, which proves (i).

To prove (ii), let $z \in C^0 - E^0$ and recall from (12) that $\rho(z) = \rho^* - r(z \to w_z) + r(D \to z)$, where $z \to w_z$ is an easy edge. We know from Lemma 3 that $r(z \to w_z) = 1 + G(S(z))$ and from Lemma 4 that $r(D \to z) = -f_1 W(z) + nf_2$. Hence

$$\forall z \in C^0 - E^0 : \rho(z) = \rho^* - 1 - G(S(z)) - f_1 W(z) + nf_2 . \tag{21}$$

By (18),

$$\rho(e) = \rho^* - 2 - f_1 W(e) + nf_2 . \tag{22}$$

Comparing (21) and (22) we see that $\rho(e) < \rho(z)$ if

$$f_1[W(z) - W(e)] < 1 - G(S(z)) . \tag{23}$$

The right-hand side of (23) is greater than ½ because we assumed that $G(S(z)) < 1/2$ for all *z* (see condition (3)). The left-hand side is the sum of *n* differences $\sum_i F(u_i) - F(u_i')$, which is at most ½ because we assumed that $0 < F(\cdot) < 1/2n$ (condition (3) again). Hence (23) holds for all nonequilibrium *z* and thus $\rho(e) < \rho(z)$.

To prove (iii) we claim that $\rho(z) < \rho(D) = \rho^*$ for all $z \in C^0$. Notice that (21) holds for *all* $z \in C^0$ because $E^0 = \varnothing$. Now observe that by condition (3),

$$-f_1 W(z) + nf_2 \le n \cdot \max_u F(u) < 1/2 \quad \text{and} \quad G(\Delta u) > 0. \tag{24}$$

From this, and (21) it follows that $\rho(z) < \rho^*$. This completes the proof of (17).

We shall now turn to the two cases of the theorem: a pure Nash equilibrium exists and a pure Nash equilibrium does not exist.

Case 1. A pure Nash equilibrium exists ($E^0 \neq \varnothing$).

By (17) we know that all welfare maximizing equilibrium states uniquely minimize the stochastic potential among all recurrence classes. Hence the states contained in

$$E^* = \{e \in E^0 : W(e) = W(e^*)\}. \tag{25}$$

constitute the stochastically stable states.

Case 2. No pure equilibrium exists ($E^0 = \varnothing$).

The chain of inequalities (17) is now reduced to (iii) since there are no equilibrium states. Hence D cannot be stochastically stable. Therefore the stochastically stable states are precisely those z that minimize

$$\begin{aligned} \rho(z) &= \rho^* - 1 - G(S(z)) - f_1 W(z) + nf_2 \\ &= g_1 S(z) - f_1 W(z) + (nf_2 + \rho^* - 1) \end{aligned} \tag{26}$$

which is equivalent to maximizing $f_1 W(z) - g_1 S(z)$. This concludes the proof of theorem 1.

6. Extensions

The proof is constructed in a way that allows us to see the impact of relaxing our assumptions on the acceptance functions F and G. In particular let us suppose that they differ among players and are not necessarily linear. Suppose that *i*'s probability of accepting the outcome of an experiment is governed by the function $G_i(\Delta u_i)$, and that *i*'s probability of stopping a search is governed by the function $F_i(u_i)$. As before we assume that these functions are strictly decreasing and satisfy

$$0 < G_i(\Delta u_i) < 1/2 \text{ and } 0 < F_i(u_i) < 1/2n \text{ for all } i. \tag{27}$$

Now let us redefine the welfare function as follows:

$$\tilde{W}(z) = -\sum_i F_i\left(u_i(\bar{a})\right). \tag{28}$$

where $\bar{a}$ denotes the vector of benchmark actions in state z.

Notice that since the functions F_i are monotone decreasing, $\tilde{W}(z)$ is monotone increasing in each player's utility $u_i(\bar{a})$. In particular $\tilde{W}(\cdot)$ is maximized at a Pareto optimal state, and is maximized among all equilibria at a Pareto optimal equilibrium.

Define:

$$S_i(z) = \max_{a_i \in A_i} \left\{u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a})\right\} \tag{29}$$

with $u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a}) > 0$. Let also

$$\begin{aligned} \tilde{S}_i(z) = -G_i\left(S_i(z)\right) &= -G_i\left(\max_{a_i \in A_i} \left\{u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a})\right\}\right) \\ &= -\min_{a_i \in A_i} \left\{G_i(u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a}))\right\} \end{aligned} \tag{30}$$

where the maximization turns into a minimization since the functions G_i are decreasing. Now redefine the stability as

$$\tilde{S}(z) = \min_i \left\{ \tilde{S}_i(z) \right\} = -\min_i \min_{a_i \in A_i} \left\{ G_i(u_i(a_i, \bar{a}_{-i}) - u_i(\bar{a})) \right\} \quad (31)$$

The logic behind the latter is that a state is more stable the bigger $\left| \tilde{S}(z) \right|$ is (that is the smaller $\tilde{S}(z)$ is), since then it is less likely that a player accepts an experiment.

A straight forward modification in the proof of theorem 1 leads to the conclusion that:

(i) If the game has a pure Nash equilibrium then every stochastically stable state is an equilibrium state that maximizes $\tilde{W}(z)$ among all equilibrium states;

(ii) If the game has no pure Nash equilibrium the stochastically stable states maximize $\tilde{W}(z) - \tilde{S}(z)$ among all $z \in C^0$.

References

Asadpour, Arash and Saberi, Amin, 2009, "On the inefficiency ratio of stable equilibria in congestion games", *5th Workshop on Internet and Networks Economics*, 545-552.

Babichenko, Yakov, 2009, "Completely uncoupled dynamics and Nash equilibria," Working Paper, Center for the Study of Rationality, Hebrew University.

Blume, Lawrence E., 1993, "The statistical mechanics of strategic interaction," *Games and Economic Behavior*, 4, 387-424.

Blume, Lawrence E., 1995, "The statistical mechanics of best-response strategy revision," *Games and Economic Behavior*, 11, 111-145.

Blume, Lawrence E., 2003, "How noise matters," *Games and Economic Behavior*, 44, 251-271.

Bowling, Michael, and Manuel Veloso, 2002, "Multi-agent learning with a variable learning rate," *Artificial Intelligence*, 136, 215-250.

Foster, Dean P., and Young, H. Peyton, 2006, "Regret testing: learning to play Nash equilibrium without knowing you have an opponent", *Theoretical Economics*, 1, 341-367.

Germano, Fabrizio, and Gabor Lugosi, 2007, "Global convergence of Foster and Young's regret testing," *Games and Economic Behavior*, 60, 135-154.

Hart, Sergiu, and Andreu Mas-Colell, 2003, "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, 93, 1830-1836.

Hart, Sergiu, and Andreu Mas-Colell, 2006, "Stochastic uncoupled dynamics and Nash equilibrium," *Games and Economic Behavior*, 57, 286-303.

Kandori, Michihiro, George Mailath, and Rafael Rob, 1993, " Learning, mutation, and long-run equilibrium in games," *Econometrica*, 61, 29-56.

Mannor, Shie, and Jeff S. Shamma, 2007, "Multi-agent learning for engineers," *Artificial Intelligence*, 171, 417-422.

Marden, Jason R., Young, H. Peyton, Arslan, Gürdal, Shamma Jeff S., 2009, "Payoff-based dynamics for multiplayer weakly acyclic games", *SIAM Journal on Control and Optimization,* 48, No. 1, 373-396.

Papadimitriou, Christos, 2001, "Algorithms, games and the internet", *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing*, 749-753.

Roughgarden, Tim, 2005, *Selfish Routing and the Price of Anarchy,* Cambridge Mass: MIT Press.

Sandholm, W. H., 2002, "Evolutionary implementation and congestion pricing, " *Review of Economic Studies*, 69, 667-689.

Shah, Devavrat and Shin, Jinwoo, 2010, "Dynamics in Congestion Games", *ACM SIGMETRICS (preliminary version).*

Thuijsman, F., Bezalel Peleg, M. Amitai, and Avi Shmida, 1995, "Automata, matching, and foraging behavior in bees," *Journal of Theoretical Biology*, 175, 305-316.

Young, H. Peyton, 1993, "The evolution of conventions", *Econometrica*, 61, 57-84.

Young, H. Peyton, 2009, "Learning by trial and error", *Games and Economic Behavior*, 65, 626-643.